# Some views on agreement to be used in content analysis studies

Popping, Roel

*Published in:*
Quality & Quantity

*DOI:*
10.1007/s11135-009-9258-3



*Publication date:*
2010

Qual Quant (2010) 44:1067–1078
DOI 10.1007/s11135-009-9258-3

# Some views on agreement to be used in content analysis studies

**Roel Popping**

Published online: 31 May 2009


**Abstract** In content analysis studies texts might be coded three times or more, certainly in the training part. With respect to reliability several views on agreement can be used now, as these are found in the literature. These views are presented, and this is followed by a discussion resulting in the opinion that in research only one view, that of mean pairwise agreement, should be used.

**Keywords** Nominal scale agreement · Interjudge reliability · Content analysis · Text analysis

## 1 Introduction

When content analysis is applied and the coding can not automatically be performed by using a computer program the investigator is confronted with a time consuming task. The text segments, which are the units of analysis, or words or themes within these segments have to be assigned to the categories that have been developed. Raters do this, where each rater usually codes a different part of the texts. First however, there is a test to evaluate how consistently raters do the coding. They all have to code the same texts, usually a subset of the complete sample, and the results of these codings are compared. As soon as the outcome of this comparison is good enough according to the investigator, each rater can proceed with a different part. This test whether the assignments resemble each other in a sufficient way is called a reliability check. Actually the purpose of reliability assessment is to assure that a data generating process can be replicated elsewhere, by other investigators, using the same coding instructions and the same text but different raters, and can therefore be relied upon in the present study.

R. Popping (✉)
Department of Sociology, University of Groningen, Grote Rozenstraat 31, 9712 TG Groningen, The Netherlands
e-mail: r.popping@rug.nl

Such a reliability check is not always performed. Lombard et al. (2002) found that only 69% (N = 137) of the articles they investigated in which a content analysis was reported contained any information on interrater reliability. In these articles even few details were provided in the average 4.5 sentences devoted to reliability and important information was ambiguously reported, not reported, or represented inappropriate decisions by researchers. Riffe and Freitag (1997) report that in 56% of the publications using content analysis in the journal Journalism Quarterly in the years 1971–1995 interrater reliability has been computed.

The reliability check is necessary. Measures are available, but there are different opinions with regard to the definition of the 'operating consistently'. Three views are found in literature. In one view there is agreement with respect to a unit when all codings are into the same category. In another view a majority rule is applied (at least k out of the m codings suffices), and in the final view the proportion of pairs of codings into the same category is used. The question is whether this number of views is really needed.

First we will go into the reliability check in context analysis. Next the general type of reliability index that is used most today for this kind of studies is presented, followed by an elaboration into the three views just mentioned. The three views are made visible in an example.

After that the question is discussed whether this large number of views is needed.

## 2 Reliability in content analysis studies

According to Holsti (1969, p. 135): "Reliability is a function of coders' skill, insight, and experience, clarity of categories and coding rules which guide their use, and the degree of ambiguity in the data." Therefore the types of indices have to be considered in the light of these three issues.

In the representational coding process (Popping 2000, p. 26), the rater uses 'Verstehen' to encode the texts according to the meanings their sources intended. The issue is not 'how' to encode text (instrumental approach), but 'whether' one chooses to apply one's own theory or one's sources' theories to the texts under analysis. Moreover the rater might even want to code implicit concepts (Carley 1994, p. 726), i.e., words or phrases that occur in the text only by implication.

Coding tasks are divided into three types (Crittenden and Hill 1971, p. 1078): A, B1, and B2. "Type A coding tasks require a coder to find a specific answer to an explicit question at a given place on an instrument. Type B1 coding tasks involve locating relevant information within a larger context ..., type B2 coding tasks are those where the coder has not only to locate relevant information, but also to evaluate the relative importance of two or more possible responses to arrive at a single code" (Montgomery and Crittenden 1977, p. 236). The type B2 includes recognizing the implicit meaning of texts as mentioned by Carley. These meanings refer to the latent content. With respect to latent content the raters must provide subjective interpretations based on their own mental schema. This "only increases the importance of making the case that the judgments of raters are intersubjective, that is, those judgments, while subjectively derived, are shared across raters, and the meaning therefore is also likely to reach out to readers of the research" (Potter and Levine-Donnerstein 1999, p. 266). Opposite to the latent meaning is the manifest meaning of words or phrases, this refers to the surface meaning of that word or phrase.

In his definition of reliability as mentioned Holsti distinguished three aspects:

- Raters' skill, insight, and experience;
- Clarity of categories and coding rules which guide their use;
- The degree of ambiguity in the data.

The first aspect refers to errors made by the raters. Such errors are reduced by better training or by using a computer-aided coding. Especially when the instrumental view is followed the computer-aided coding is very reliable. In training programs attention should be given to the following (Popping 2000, p. 139):

- Raters should be aware of the goal of the investigation;
- Raters should exactly know the meaning of the categories used (also part Holsti's second aspect);
- Raters should be trained in identifying the target behavior. In case they have to code different categories at once, they should be trained in which category to look first at;
- In case there are reasons to assign different categories to a text unit, they should know which category is to be preferred in which situation, and why.

If categories or coding-rules are not clear one might be confronted with procedureerrors. Therefore the rater must know which words or phrases refer to which specific category. The decision rules in complex situations must also be clear. In the pilot phase of a study, the researcher might investigate the frequency with which different categories are used for assigning specific units, i.e., he looks at the non-diagonal cells of the agreement table. A high frequency in such a cell might indicate that the disagreement perhaps is not due to the raters, but to indistinctness of the categories. One way to solve such a problem is adapting or merging the categories.

The coding process can be structured for human raters with the aid of computer programs that guides the correct application of coding rules. Now reliability will increase as words or phrases will only be coded in existing categories.

Ambiguity of language is at present at many places and in different forms (Popping 2000, p. 21). Investigators have to be aware of it; computer-aided methods can assist in finding the ambiguity, especially when the representational view is followed. There are no tools (yet), like expert systems, to validly explicate linguistic relations in texts.

## 3 Reliability indices

The kind of reliability that is considered here is denoted as agreement. There is a difference however between reliability and agreement. This is well described by Tinsley and Weiss (2000, p. 98): "Interrater reliability provides an indication of the extend to which the variance in the ratings is attributable to differences among the objects rated… Interrater agreement represents the extent to which the different judges tend to assign exactly the same rating to each object." Concerning the level of measurement they notice however that "[t]he distinction between interrater reliability and interrater agreement ceases to exist when ratings are made at the nominal level of measurement" (Tinsley and Weiss 2000, p. 101).

Agreement is also considered as a special kind of association. Again there is a difference. It is important to determine the similarity of the content of behavior (in a broad sense) between raters in general with the degree of identity of this behavior. The behavior of one rater does not have to be predicted from that of the other. In the case of association one investigates the strength of the linear relationship between variables. Here the goal is to predict the values

of one variable from those of the other. With regard to agreement, most important is the similarity of the content of behavior between raters, with the goal of determining the degree of identity of this behavior (see Popping (1985), about the same reasoning is followed by Krippendorff (1987)). The basic idea of an agreement index is looking at the proportion of units on which raters agree in assigning to a category.

Consensus is growing among investigators who use indices for nominal scale agreement in their research that the preferred index should be of the type

$$I = \frac{O - E}{M - E},$$

where O stands for observed, E for expected, and M for maximum agreement (Galtung 1979). The index also has to suffice several quality criteria (Popping 1988). Two very important ones are that the maximum value of the index should be 1, no matter the number of raters, units of analysis or categories, and that in case of statistical independence between the assignments by the raters the index has to take the value 0. Even in this extreme case of independence the raters will agree on the classification of some units of analysis, purely by chance. The correction for such chance agreement has played an important role in the literature.

Several agreement indices of the above type are known. They differ in their definition of chance agreement. Scott (1955) has proposed an index for comparing the assignments by two raters based on the idea that an equality relation exists between these raters, they are drawn from one universe. Therefore the computation of the expected agreement should be based on one stochastic variable, referring to the mean distribution of the categories across all raters. He proposed an index, called pi, $\pi$. This view is criticized by Cohen, "one source of disagreement between a pair of judges is precisely their proclivity to distribute their judgments different over the categories" (Cohen 1960, p. 41). Therefore for each rater a stochastic variable exists, containing the distribution over the categories according to that rater. Cohen now proposed a slightly different index, called kappa, $\kappa$. Confusion increased when Fleiss (1971) proposed an extension to the situation where assignments by more than two raters were pairwise compared. Fleiss based himself on Cohen, denoted the index also as kappa, but followed Scott in the computation of chance expected agreement.

The different starting positions between these two authors can easily be understood from their field of application. Scott, doing text analysis, needed many raters to analyze the answers on open-ended questions used in survey research. He was just interested in 'check-coding': if one part of the data was judged by one rater and another part by another rater, would that make a difference? Cohen, in the field of clinical psychology, had all units classified by all raters. These raters (often psychologists, psychiatrists or physicians) had their qualifications that should be taken into account. In case text analysis is applied the view as initially expressed by Scott is the one generally to be preferred.[1] Measures of the type intended are often indicated by the name kappa. This is followed by mentioning the specific research situation for which the index is meant.

[1] The development in terms of history of the kappa-like index is as follows. At first the percentage of cases in which both raters agree in classifying units of analysis was used as an agreement index. Rogot and Goldberg (1966) denote this proportion as the "index of crude agreement". The disadvantage of this index, however, is that in cases with few categories the probability of equal assignments is greater than in cases with many categories, especially when one takes into account the agreement that might be expected by chance. Therefore Bennett et al. (1954) proposed a correction for the number of categories. They assume that all categories have equal probability to be used. An index in which this often unrealistic assumption is removed was the already mentioned one by Scott. Also is already mentioned the critics by Cohen on Scott. Note however, that they referred to different research situations.

From here on the line proposed by Scott will be followed, for this line fits best to the situation of content analysis. It means that expected agreement is based on one stochastic variable, the mean distribution of the categories across all raters. In many studies only two raters are used. It would be better to formulate that coding is performed only twice, not necessarily by the same raters. This is a minimal requirement. To assure replicability one needs a sufficiently large number of raters. If the qualifications of raters are widely shared, then the investigator has to assure that raters are chosen having the needed qualifications. In practice this is difficult.

In case the same units of analysis have been classified more than twice however, some more care is required in defining the type of agreement one is looking for. Several views on agreement are possible. These are denoted as *mean pairwise*, *simultaneous*, and *majority* agreement. In the first situation observed and expected agreement are first computed across units of analysis averaged over all codings or, in case of fixed raters, for each pair of raters. The means are entered into the formula for kappa. One speaks about simultaneous or strict agreement in case all raters have assigned a unit of analysis to the same category. Majority agreement contains a weaker view, this type of agreement holds if and only if at least k of the m codings per unit of analysis have been assigned to the same category (here k is chosen by the investigator)[2].

It is my guess that the views came up in the heads of those developing the formulas, and that there have been no clear thoughts on what such a view might hold. Let us look at the two articles where the indices for mean pairwise agreement have been introduced. In the first one it is stated: "The extension follows directly from the idea of taking pairs of observers' assignment of an item, and seeing whether both observers assign the item to the same category or to different categories" (Light 1971, p. 368). The other one reports: "In this paper we consider only the generalization of unweighted kappa to the measurement of agreement among any constant number of raters where there is no connection between the raters judging the various subjects" (Fleiss 1971, p. 378). The motivation was found in the technical possibility.

In this paper the three views on agreement to be used in the situation content analysis is applied, is elaborated, and the different indices are presented. The question why so many views might be necessary is also discussed. These views have only sense in case the units of analysis have been coded more than twice. In case all units have only been coded twice, pairwise and simultaneous agreement become identical and majority agreement is not possible.

## 4 Three views on agreement

Three views on agreement will be presented; these are mean pairwise agreement, simultaneous agreement, and majority agreement. The situation in which units of analysis are only judged twice is a special case of both pairwise and simultaneous agreement. Therefore the situation will not be discussed separately. Majority agreement is only possible in case of at least three judgments.

The general formulation for the kappa agreement index is

$$K = \frac{P_o - P_e}{1 - P_e}, \tag{1}$$

[2] In case k = m simultaneous and majority agreement are identical. I do not go into the discussion whether simultaneous agreement is a special case of majority agreement or whether majority agreement is a weaker version of simultaneous agreement.

where $P_o$ denotes the observed proportion of agreement, and $P_e$ the proportion of agreement under the null-hypothesis of independence. In the different views on agreement only the ways $P_o$ and $P_e$ are defined differ.

Hereafter it is assumed that there are N units of analysis to be assigned to c categories. Each unit of analysis is judged m times. The number of times unit of analysis s is assigned to category i is denoted by $n_{si}$. The formulas presented are, as indicated before, based on the view by Scott.

### 4.1 Pairwise agreement

The agreement index as developed by Scott has first been extended into the direction of average pairwise agreement among more than two codings. The first two papers on this topic appeared at the same moment (Fleiss 1971; Light 1971). It took some years until Conger (1980) explained the difference between the indices proposed in these papers. In his article Light followed Cohen in defining expected agreement. Actually Fleiss followed Scott in the definition of expected agreement, but gave the impression he followed Cohen. Define

$$P_{ii} = \sum_{s=1}^{N} n_{si}\,(n_{si} - 1)/(Nm\,(m-1)) \tag{2}$$

as the proportion of observed agreement with regard to category i, and

$$P_i = \sum_{s=1}^{N} n_{si}/(Nm) \tag{3}$$

as the fraction with which category i is used. All $P_i$ sum to 1. The observed proportion of agreement now is

$$P_o = \sum_{i=1}^{c} P_{ii} \tag{4}$$

and the expected proportion

$$P_e = \sum_{i=1}^{c} P_i^2. \tag{5}$$

$P_o$ and $P_e$ are substituted in the formula for kappa (1). $P_e$ is based on the mean proportion with which each category is used by the raters. The distribution of the categories per rater is not used.

In case there are fixed raters, one can also compute the mean of the kappas across all pairs of raters. The kappa based on mean observed and expected agreement is a weighted mean of the separate kappas. In this kappa extreme high or low amounts of expected agreement between a pair of raters are entered into the expected mean in this proportion.

### 4.2 Simultaneous agreement

Simultaneous agreement refers to the situation in which it is decided that there is only agreement within the judgments with respect to a unit of analysis in case in *all* assignments the unit is assigned to the same category. Hubert (1977, p. 296) denoted this type of agreement as De Moivre's definition of agreement.

Simultaneous agreement looks especially useful in case the investigators demands are extremely high. Just one deviating judgment of a unit of analysis satisfies to decide that there is no agreement with respect to that unit. Define

$$a_s = \sum_{i=1}^{c} a_{si}, \tag{6}$$

where $a_{si} = 1$ in case $n_{si} = m$, otherwise $a_{si} = 0$. Now

$$P_o = \sum_{s=1}^{N} a_s / N \tag{7}$$

is the proportion observed agreement within the judgments. The proportion of expected agreement is defined as

$$P_e = \sum_{i=1}^{c} P_i^m. \tag{8}$$

The results found in (7) and (8) have to be substituted in the formula for kappa (1). Ross (1977) and Hubert (1977) defined the index following Cohens view. Many authors paid attention to the index.

### 4.3 Majority agreement

In majority agreement there is already a contribution to agreement in case at least k out of the m judgments of a unit of analysis are to one and the same category. For the computation of the amount of observed agreement formulas (6) and (7) can be used, with the difference that now $a_{si} = 1$ in case $n_{si} \geq k$, and otherwise again $a_{si} = 0$. A condition with regard to k is that $k \geq (m + c - 1) / c$, and $k > m/2$, otherwise $a_{si}$ will by definition take the value 1, which means that one will always find that there is perfect agreement among the judgments. As an example consider the situation in which $k = 2$, $m = 5$, and $c = 4$. Because there are 5 judgments and 4 categories, at least two units have to be assigned to the same category. This satisfies the requirement $k = 2$, so there is agreement with regard to the unit of analysis. This however, holds for any unit of analysis. Therefore the condition is necessary.

The amount of expected agreement is

$$P_e = \frac{1}{c} \sum_{r=k}^{m} \sum_{i=1}^{c} \binom{m}{r} p_i^r \left(1 - p_i\right)^{m-r}. \tag{9}$$

The values for $P_o$ and $P_e$ should again be substituted in formula (1).

Conger (1980) has presented a formula for the computation of agreement in case just k out of the m judgments per unit of analysis are into the same category, and Craig (1981) presented a formula for the 2 out of 3 codings situation.

## 5 Extensions

All indices discussed have been extended. Versions are available in which weights are used, which makes the indices appropriate for data at an ordinal or interval level of measurement. There is a version of the kappa for mean pairwise agreement in which it is allowed that all units of analysis are not judged the same number of times. Intra and interclass versions of

**Table 1** Agreement frequency per category for data that are judged four times (hypothetical data), and their contributions to the observed agreement

| Category | | | | Contribution to $P_o$ | | |
|---|---|---|---|---|---|---|
| # Unit | a | b | c | Pairwise | Majority | Simultaneous |
| 01 | 4 | – | – | 1 | 1 | 1 |
| 02 | – | 1 | 3 | .5 | 1 | – |
| 03 | 2 | 1 | 1 | .17 | – | – |
| 04 | 1 | 3 | – | .5 | 1 | – |
| 05 | 2 | – | 2 | .33 | – | – |
| 06 | – | 4 | – | 1 | 1 | 1 |
| 07 | 3 | 1 | – | .5 | 1 | – |
| 08 | 1 | 2 | 1 | .17 | – | – |
| 09 | 3 | – | 1 | .5 | 1 | – |
| 10 | – | 4 | – | 1 | 1 | 1 |
| 11 | 1 | – | 3 | .5 | 1 | – |
| 12 | 3 | 1 | – | .5 | 1 | – |
| 13 | – | – | 4 | 1 | 1 | 1 |
| 14 | – | 4 | – | 1 | 1 | 1 |
| 15 | – | 1 | 3 | .5 | 1 | – |
| 16 | – | 2 | 2 | .33 | – | – |
| 17 | 3 | 1 | – | .5 | 1 | – |
| 18 | – | 4 | – | 1 | 1 | 1 |
| 19 | 1 | – | 3 | .5 | 1 | – |
| 20 | – | 1 | 3 | .5 | 1 | – |
| Sum | 24 | 30 | 26 | 12 | 16 | 6 |

the indices have been developed. Sampling characteristics of all these types of indices are also known.

For the situation in which the view of Cohen is followed it is also possible to compare the raters to a standard and to compute the best out of a series of raters. For further details on all these extensions, the reader is referred to Popping (1992).

## 6 Example

As an example take the data from Table 1. Twenty units of analysis have been judged four times, they have been assigned to one of three possible categories. The data are hypothetical. For majority agreement, there is a contribution to agreement in case at least 3 out of the 4 judgments are equal.

In the right part of the table the contribution of each unit of analysis to the observed agreement according to the methods of pairwise, majority, and simultaneous agreement is presented.[3]

[3] All computations have been performed by using the computer program Agree 7.3 (Popping 2000).

**Table 2** Kappas for hypothetical data (N = 20)

| | $\kappa$ | $P_o$ | $P_e$ |
|---|---|---|---|
| Pairwise agreement | .40 | .60 | .34 |
| Majority agreement | .77 | .80 | .11 |
| Simultaneous agreement | .27 | .30 | .04 |

The values that are computed for the indices based on the three methods are given in Table 2.

From these values one can see that the most severe definition of agreement, simultaneous agreement, results in the lowest value for kappa. The highest value is found in the situation of majority agreement. However, it is not necessarily so that majority agreement should be greater than pairwise agreement. This depends on whether it occurs more that there is at least agreement on 3 out of the 4 judgments per unit of analysis, or not. If not, there is no contribution to majority agreement, but there is to pairwise agreement. There is always agreement on at least 2 out of the 4 judgments.

The agreement score usually is not the end of the research. The investigator wants to confront the item that is coded with other items. Therefore one score on the item is needed for each unit of analysis. It is relevant for the investigator to know that it does not matter which rater assigned the score, or that in case another rater had performed the coding (nearly) the same classification would have resulted. The value the agreement index takes should meet some criterion. Generally this criterion is a rule of thumb as mentioned in literature, e.g. Landis and Koch (1977, p. 165) claim there is almost perfect agreement in case $\kappa \geq .81$. Here no attention is given to aspects of the actual investigation like for instance the difficulty of the coding task. This is to be regretted.

As indicated, usually the assignments by one of the raters are used in the further analyses. It is possible to assign each unit of analysis to a category based on the assignments by all raters. When agreement is found for at least 3 out of the 4 judgments the corresponding category is used. This also holds in the situation of agreement on two scores, while the other two are in different categories. In case there is agreement on two categories but also on two other categories or when all assignments are different, one can only take one score at random. In case however, the categories are on an ordinal scale, the median of the judgments sometimes can be used.

## 7 When to use what

It has been shown that indices for pairwise, simultaneous and majority agreement are available. From a mathematical perspective they all can be defined. However, there has not been a discussion about the question why exactly the index as presented is needed. Now one can at least focus on the way an index behaves when agreement is increasing or decreasing, and one can look at what is really measured.

### 7.1 Behaviour of the index

First, we might look at how the indices behave when the number of differences in coding increases. Figure 1 contains data on 30 units that are each coded three times into one of 3 categories. Each rater used each category 10 times. Therefore the expected agreement is for

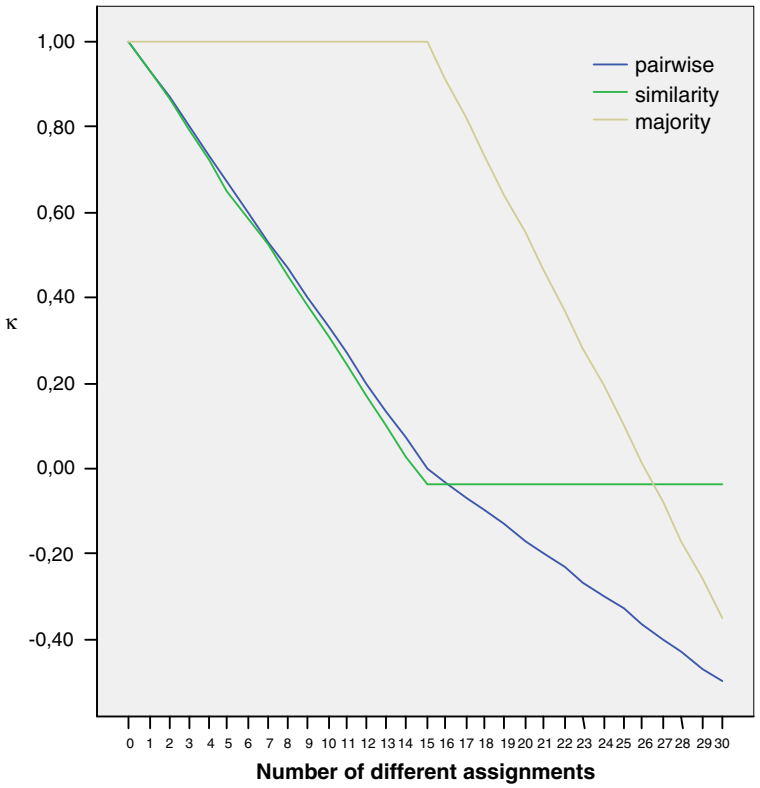


**Fig. 1** Kappa values for the three methods (n = 30)

each method in all situations identical; i.e., for pairwise, simultaneous and majority agreement .33, .04 and .26. Differences in the outcome of the index are caused by the observed agreement. At first there was no disagreement in coding, for this reason kappa will take the value 1. In the second step one rater had once assigned a unit to category 2 instead of category 1 and once to category 1 instead of category 2. In the situations of pairwise and simultaneous agreement kappa will now take the value .93. Note however that observed agreement is different. In the pairwise comparison it is .96, in the simultaneous comparison .93 and in the case of majority agreement the value is still 1. In this situation the agreement index can take a high value very soon. In the worst case all raters disagree in assigning the categories. They all use a different category. Now the kappa for pairwise agreement is −.50, for similarity agreement −.04 and for majority agreement −.33.

As long as at least 2 out of the 3 codings are in agreement the kappa for majority agreement takes the value 1. Here, this is in half of the situations. In the last one of these situations it holds for all units that only one pair is in agreement. Now the kappa for pairwise agreement will take the value 0. As soon as the codings are no longer in agreement for at least one unit, the observed agreement for the simultaneous version of the index becomes 0, and therefore the kappa takes a constant value −.04. This also happens in half of the situations, but here we have a kappa nobody should be satisfied with. Only the kappa for mean pairwise agreement steadily decreases from 1 to −.50. This might be considered as a plea for the pairwise $\kappa$.

### 7.2 What is measured

In his description of what is expressed in reliability Holsti referred to the degree of ambiguity in the texts. This concerns in part the complexity of the coding task (type B2), it also concerns the skills the raters have. Also the clarity of the categories was mentioned. The results of an agreement study might say something about this last issue.[4] Reliability also gives insight in the skills of the raters. These skills arise from the level and quality of the training and from the experience a rater has.

These skills can be linked to the way agreement is defined. The value the index will take expresses how well the raters in general did their job, i.e., to what extend they came to identical assignments (note again, this does not by definition mean it is the correct assignment.).

If, on the level of a unit, one coding is into a different category, there is no contribution to simultaneous agreement. In this situation one is 'punished' because of one deviating assignment. When there are sufficient categories, however, it is also possible that all codings were to a different category. These two possibilities are not distinguished, so whether there is agreement in the other assignments does not count. In order to contribute to agreement the demands are extremely high. All codings must be into the same category.

This demand might be too high. Now majority agreement comes in. Here the demand is not that high. Usually the definition is taken that only at most one coding is allowed to be in another category. Here too however, in case this does not hold, the investigator does not know whether only two codings are into a different category (this might even be the identical category), or all codings are into different categories. Here too, the correspondence in the remaining assignments is not counted.

In the situation of pairwise agreement, all identical pairs are counted. So when 5 codings have been performed there are 10 pairs. Say, 3 of these are into one specific category, and the other 2 into another category. Now .4 of all pairs is in agreement. In case 4 codings would have been into the same category, the contribution to agreement would be .6. In the situation of simultaneous agreement both contributions would have been 0, and in the situation of majority agreement (3 out of 5) they would have been 1.

It looks like mean pairwise agreement contributes to the final statistic in a way that is clearly understood, and it informs in general on the quality of all assignments.

## 8 Conclusion

In this contribution we have been looking for arguments why one of the types pairwise, simultaneous or majority agreement is to be preferred over the others in case one is looking for agreement between codings when the raters are assumed to be equally skilled. Looking at the development of the outcome values in a situation where agreement decreases in a systematic way mean pairwise agreement looks preferable. In case one looks at the contribution of the assignments to one single unit pairwise agreement tells best what is judged. Therefore it is preferred that the mean pairwise agreement is used in all studies using more that two codings per unit. Now this already happens in a considerable part of the studies, but it should in all studies.

[4] One might investigate the agreement per category (Schouten 1980). A low amount of agreement indicates it was hard for the raters to decide on assigning an entity to the category under investigation or to some other category. This might have to do with the fact that the category is not clear. For the data from Table 1 we find for the three categories: $\kappa = .37, .43, .39$, which might indicate that the raters had most difficulties with assigning to the first category. The differences however are small.

## References


Bennett, E.M., Alpert, R., Goldstein, A.C.: Communications through limited response questioning. Public Opin. Q. **18**(3), 303–308 (1954)

Carley, K.M.: Content analysis. In: Asher, R.E. (ed.) The Encyclopedia of Language and Linguistics, pp. 725–730. Pergamon Press, Edinburgh (1994)

Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**(1), 37–46 (1960)

Conger, A.J.: Integration and generalization of kappas for multiple raters. Psychol. Bull. **88**(2), 322–328 (1980)

Craig, R.T.: Generalization of Scott's index of intercoder agreement. Public Opin. Q. **45**(2), 260–264 (1981)

Crittenden, K.S., Hill, R.J.: Coding reliability and validity of interview data. Am. Sociol. Rev. **36**(6), 1073–1080 (1971)

Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**(5), 378–382 (1971)

Galtung, J.: Measurement of agreement. In: Galtung, J. (ed.) Papers on Methodology. Theory and Methods of Social Research, Vol. II, pp. 82–135. Christian Eijlers, Copenhagen (1979)

Holsti, O.R.: Content Analysis for the Social Sciences and Humanities. Vol. 2, Addison Wesley, London (1969)

Hubert, L.J.: Kappa revisited. Psychol. Bull. **84**(2), 289–297 (1977)

Krippendorff, K.: Association, agreement, and equity. Qual. Quant. **21**(2), 109–123 (1987)

Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)

Light, R.J.: Measures of response agreement for qualitative data: some generalizations and alternatives. Psychol. Bull. **76**(5), 365–377 (1971)

Lombard, M., Snyder-Duch, J., Bracken, C.C.: Content analysis in mass communication: assessment and reporting of intercoder reliability. Hum. Commun. Res. **28**(4), 587–604 (2002)

Montgomery, A.C., Crittenden, K.S.: Improving coding reliability for openended questions. Public Opin. Q. **41**(2), 235–243 (1977)

Popping, R.: Nominal scale agreement. In: Kotz, S., Johnson, N.L. (eds.) Encyclopedia of Statistical Sciences, pp. 261–264. Wiley, New York (1985)

Popping, R.: On agreement indices for nominal data. In: Saris, W.E., Gallhofer, I.N. (eds.) Sociometric Research, pp. 90–105. McMillan, London (1988)

Popping, R.: Taxonomy on Nominal Scale Agreement. Vol. I, iec *Pro*Gamma, Groningen (1992)

Popping, R.: Computer-assisted Text Analysis. Sage, London (2000)

Potter, W.J., Levine-Donnerstein, D.: Rethinking validity and reliability in content analysis. J. Appl. Commun. Res. **27**(3), 258–284 (1999)

Riffe, D., Freitag, A.A.: A content analysis of content analyses: twenty-five years of journalism quarterly. Journal. Mass Commun. Q. **74**(4), 873–882 (1997)

Rogot, E., Goldberg, I.D.: A proposed index for measuring agreement in test–retest studies. J. Chronic Dis. **19**(9), 991–1006 (1966)

Ross, D.C.: Testing patterned hypothesis in multiway contingency tables using weighted kappa and weighted chi square. Educ. Psychol. Meas. **37**(2), 291–307 (1977)

Schouten, H.J.A.: Measuring pairwise agreement among many observers. Biom. J. **22**(6), 497–504 (1980)

Scott, W.A.: Reliability of content analysis: the case of nominal scale coding. Public Opin. Q. **19**(3), 321–325 (1955)

Tinsley, H.E.A., Weiss, D.J.: Interrater reliability and agreement. In: Tinsley, H.E.A., Brown, S.D. (eds.) Handbook of Applied Multivariate Statistics and Mathematical Modelling, pp. 95–124. Academic Press, San Diego (2000)